College Football Statistics

Cody A Faircloth

Dev10

2025-3 Data Engineering

Corbin March

25 April, 2025

College Football Statistics

The College Football Statistics project is an exercise in data engineering consisting of data extraction, transformation, and loading (ETL) and data visualization. The goal of this project is to create a dashboard on par with ESPN to display college football statistics. Additionally, the dashboard can show distribution of statistics among players, teams, conferences, statistical changes over time, and the frequency of passing versus rushing throughout college football.

Technologies Used

The main programming language used in this project is Python. The relational database management system used was MySQL. Docker was used as a containerization platform with Alpine, Apache Airflow, and Apache Spark images being used. Several Python libraries were used including Airflow, Dash, DynaConf, SQLAlchemy, SSHOperator, Pandas, Pendulum, Plotly Express, and PySpark.

Approach: ETL

To effectively store data for practical use with a complex dataset like college football statistics, a complex schema with a dynamic ETL pipeline had to be designed. The finalized database schema is displayed at the following link:
https://github.com/codyfaircloth55/CFB_Statistics/blob/main/supporting-documents/cfb_erd.png

The approach for the ETL pipeline consisted of separating concerns and creating scalable functions that perform ETL for a specific table. PySpark was used in the ETL process and performed very well with the relatively large dataset provided.

Outcome: ETL

The ETL process was successful. The design I built was scalable allowing for as many years of data as loaded to be accurately processed. For this project, however, I only included five years of data spanning from 2020 to 2025. The SQL schema and Spark Script were linked to Airflow allowing for automated task scheduling.

Approach: Visualizations

The visualization approach was incredibly simple. Split the dashboard into a by team, by conference, and national section then display standings, statistics, and visualizations. This is done through distinct callbacks using specific SQL queries to retrieve the necessary data.

Outcome: Visualizations

The dashboard works well and provides accurate tables and graphs for each given input. Some needed queries were difficult to generate, but with testing on MySQL Workbench, all necessary queries were created and implemented into its respective callback function.

Conclusion

It is very possible to make a dashboard that displays statistics for single sport up to par with ESPN. The only constraint is how much data you're willing to load into the database. Additionally, there is a clear trend of more overall passing than rushing throughout college football.

References

- Datasets
    - https://www.sports-reference.com/cfb/
- Logos
    - https://www.wikipedia.org/